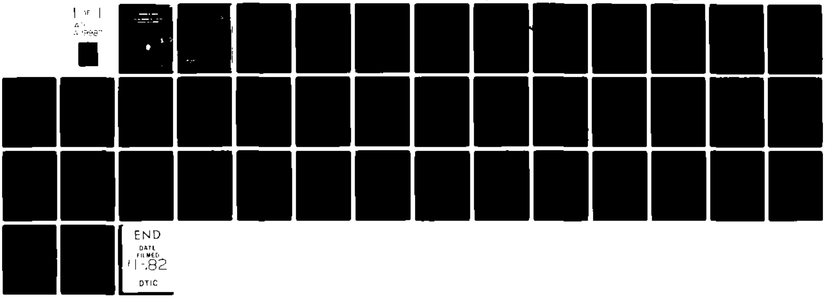
AD-A119 987 SOUTHEASTERN CENTER FOR ELECTRICAL ENGINEERING EDUCAT--ETC F/G 6/5
DATA BASE MANAGEMENT FOR U.S. AIR FORCE OCCUPATIONAL HEALTH PRO--ETC(U)
AUG 82 M S HILBERT, M CHAMPION, R G SMITH F33615-78-D-0617

UNCLASSIFIED SAM-REVIEW-4-82 NL

1 OF 1
AD
A119987



END
DATE
FILMED
11-82
DTIC

Review 4-82

AEROMEDICAL REVIEW

DATA BASE MANAGEMENT FOR U.S. AIR FORCE OCCUPATIONAL HEALTH PROGRAMS

Morton S. Hilbert, M.P.H.
Ralph G. Smith, Ph.D.
Lawrence J. Fine, M.D.
Michael Champion, M.A.
Department of Environmental and Industrial Health
School of Public Health
University of Michigan
Ann Arbor, Michigan 48109

August 1982

Approved for public release; distribution unlimited.

Prepared for USAF SCHOOL OF AEROSPACE MEDICINE
Aerospace Medical Division (AFSC)
Brooks Air Force Base, Texas 78235

Southeastern Center for Electrical Engineering Education
1101 Massachusetts
St. Cloud, Florida 32769

NOTICES

This final report was submitted by the Department of Environmental and Industrial Health, School of Public Health, University of Michigan, Ann Arbor, Michigan, under contract number F33615-78-D-0617, task number 49, job order 7757-80-01, with the USAF School of Aerospace Medicine and the USAF Occupational and Environmental Health Laboratory, Aerospace Medical Division, AFSC, Brooks Air Force Base, Texas. Dr. Michael G. Yochmowitz (USAFSAM/BRP) and Colonel Beverly J. Dye (USAF OEHL/EC) were the Laboratory Project Scientists-in-Charge.

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This report is based on research partially funded by the Laboratory Director's Independent Research (LDIR) Program.

This report has been reviewed by the Office of Public Affairs (PA) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

Michael G. Yochmowitz
MICHAEL G. YOCHMOWITZ, Ph.D.
Project Scientist

JOHN E. PICKERING, M.S.
Chief, Radiation Sciences Division

Beverly J. Dye
BEVERLY J. DYE
Colonel, USAF, NC
Project Scientist

ARTHUR P. CALDWELL
Colonel, USAF, BSC
Chief, Consultant Services Division

ROY L. DEHART
Colonel, USAF, MC
Commander

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
Aeromedical Review 4-82

2. GOVT ACCESSION NO.
AD-A119987

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
DATA BASE MANAGEMENT FOR U.S. AIR FORCE OCCUPATIONAL HEALTH PROGRAMS

5. TYPE OF REPORT & PERIOD COVERED
Final Report
August 1980 - January 1982

6. PERFORMING ORG. REPORT NUMBER
SAM-TR-82-17

7. AUTHOR(s)
Morton S. Hilbert, M.P.H. Michael Champion, M.A.
Ralph G. Smith, Ph.D.
Lawrence J. Fine, M.D.

8. CONTRACT OR GRANT NUMBER(s)
F33615-78-D-0617
Task 49

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Environmental and Industrial Health
School of Public Health
University of Michigan
Ann Arbor, Michigan 48109

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
61101F
7757-80-01

11. CONTROLLING OFFICE NAME AND ADDRESS
USAF School of Aerospace Medicine (RZ)* and
USAF Occupational and Environmental Health
Laboratory (ECO)*

12. REPORT DATE
August 1982

13. NUMBER OF PAGES
41

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)
Southeastern Center for Electrical Engineering
Education
1101 Massachusetts
St. Cloud, Florida 32769

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

*Aerospace Medical Division (AFSC)
Brooks Air Force Base, Texas 78235

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)
Occupational Health
Environmental Health
Data Base Management

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The objective of this study has been to explore the range of feasible designs for an occupational health data system which the study group deems most appropriate to meet the perceived needs of the U.S. Air Force Occupational Health Program. The main issue is that of centralization vs. decentralization. Recommendations are made for three basic configurations, and 5-year costs of storing and retrieving the base-level data are estimated. A distributed system is found to be the most desirable.

PREFACE

Passage of the Occupational Safety and Health Act of 1970 has had expanding impact on U.S. Air Force occupational safety and health programs. Federal agencies are now required to comply with the same standards as industry as a result of EO 12196, Occupational Safety and Health Programs for Federal Employees.

Tasking from USAF Research and Bioenvironmental Functions requested the USAF Occupational and Environmental Health Laboratory to recommend an approach for a new, more comprehensive occupational health program. The present report contains the results of an examination, by a non-Air-Force group of experts, of the Air Force's plans for automation of an occupational health data base.

Michael G. Yochmowitz
MICHAEL G. YOCHMOWITZ, Ph.D.
Mathematical Statistician

Beverly J. Dye
BEVERLY J. DYE, Colonel, USAF, NC
Chief, Occupational Health Branch

TABLE OF CONTENTS

DATA BASE MANAGEMENT FOR U.S. AIR FORCE OCCUPATIONAL HEALTH PROGRAMS

CHAPTER 1: SYSTEM REQUIREMENTS

Introduction

This report on Data Base Management for the USAF Occupational Health Programs has been prepared as part III of contract #F33615-78-D-0617 between The Southeastern Center for Electrical Engineering Education and The University of Michigan.

The University of Michigan study group was asked to make recommendations on 11 issues concerning the design of an occupational health data system:

a) Interactive vs batch processing

b) Central vs MAJCOM vs base data entry and processing

c) Disk vs tape storage

d) Canned programs vs custom software

e) Telephone vs hard-wire vs AUTOVON telecommunications

f) In-house vs contract programmers

g) Turnkey user system vs programmers-only system

h) Number and kinds of computer terminals

i) Single system for both civilian and military vs separate systems

j) User access vs confidentiality

k) Computer system on which occupational health software should run

The objective of this study has been to explore the range of feasible designs for an occupational health data system and to make recommendations which the group believes to be most appropriate to meet the perceived needs of the USAF Occupational Health Program. The strength and weakness of various options have been presented.

Ideal Specifications for an Occupational Health Data System

The Air Force has a rather diverse group of potential users of an occupational health data system, and specifying the requirements of a system that will adequately meet the needs of all potential user groups is not an easy task. A useful system must be fairly general, but attempting to design a system that satisfies everyone has real dangers. Experience in the computer industry tends to show that such efforts take too long to design, are very difficult to implement, and often end up doing too little for everyone rather than just enough for most.

Based on the experience of this study group, that of industry, and communications with USAF personnel, the following requirements are both necessary and feasible:

Data Collection

It is very important that epidemiologists be able to make studies relating exposures to particular chemicals to specific diseases as well as to abnormal test results. All industries visited by the study group make considerable effort to (1) collect data on the nature of the illnesses that lead to absences from work of greater than a few days, and (2) collect and code information from the death certificates of employees and retirees. The nature of the illness (or cause of death) is recorded by International Classification of Diseases (ICD) codes, and these can be related to work history and exposure data via the employee's social security number. Implementation of such a program for both military and civilian employees would pose considerable practical difficulties. Industry is, however, dealing with similar problems, often with considerable success. The Air Force should work toward a future data collection system that integrates military and civilian employees and contains data on both "occupational" diseases and those observed as part of a more general preventative medicine program. For the time being, however, a high priority should be assigned to ensuring that medical conditions observed as part of the Standardized Occupational Health Program (SOHP) are appropriately coded for computerized data retrieval and analysis.

In visits to several bases the study group noted that excessive effort is devoted to preparing diagrams of each workplace in the shop folders each time an inspection is made. This information could be computerized and then updated only when physical modifications are made to a workplace. When a graphic display is desired, the computer could generate a diagram on a high-resolution graphics terminal or plotter or by using conventional printer symbols at lower resolution. (This function is not by itself sufficiently important to justify the purchase of high-resolution graphics equipment.) In either case, the basic information about the layout of the workplace and industrial hygiene measurements could be stored in a manner that would reduce the "busy work" demanded by the current system.

Coding Improvements

In many cases, data can be stored more efficiently in computers by developing a scheme that assigns brief numeric codes representing more descriptive words or phrases. The coded information can be stored in much less memory than the descriptive information, and coding simplifies the program's operation since it does not have to search for synonyms, correct misspellings, etc., when retrieving information by the descriptor. A disadvantage of coding is that humans must generally do more work to increase the computer's efficiency; i.e., to somehow translate the description into the computer code. (The data entry specifications discussed later allow simplification of this task.) Any coding scheme drastically limits the specificity of the information entered, and care must be taken early to ensure that the code covers the range of possibilities so that one code offers at least a reasonable categorization of the information being entered. On the other hand, information must be categorized and abstracted before any reasonable analysis can take place; coding schemes force this to be done before the data are collected.

For example, rather than identifying monitoring equipment by long descriptions and more or less meaningless serial numbers, a coding scheme should be devised so that a particular instrument could be identified by a short code; then given the short code, the computer would be able to print an extended description on demand. This procedure can also be applied to hazard-abatement controls, chemical handling guidelines, routine medical tests, and anywhere else that a single symbol can be made to stand for a larger amount of standardized information.

More specifically:

a) Codes should be devised for general types of hazard controls, not for specific instances of these controls. For example, ventilation systems can be classified as to whether they are of the horizontal, vertical, downdraft, etc., types; information that would identify a unique installation is not needed. Similarly, respirators should be coded by general type (e.g., organic vapor), not unique serial numbers.

b) Workplace identification codes should be unique, but should be organized by the type of process rather than by the location of the workplace. The definition of a process (and hence a workplace) should be based on the hazards likely to be encountered and the controls used to abate them. The objective here would be to increase the plausibility of the assumption inherent in most monitoring schemes that individuals in a given workplace are exposed to similar hazards and use similar controls, so hazards and controls can be monitored by workplace and be valid estimates for the individuals who work there.

c) For most purposes, the codes or keys used to identify something or someone should be as common and obvious as possible, and established Air Force wide rather than by specific base or workplace. Occasionally privacy considerations (discussed later) will warrant that keys (such as an individual's social security number) be scrambled to inhibit unauthorized access to data, but such instances are likely to be the exception for the USAF Occupational Health Program data.

Data Entry

Accurate and efficient entry of data into the computer has posed problems for virtually all data systems. It is widely accepted that the best time to trap and correct errors is when they are first typed, especially if the person entering the data is familiar with the way in which the data were obtained and is not simply a keypuncher. That is, if the BEE (bioenvironmental engineer) or technician who makes an observation enters it into the computer and is informed that the value entered is logically impossible or highly implausible, he or she can evaluate the likely source of the error and respond appropriately and quickly. If the data are transcribed onto paper forms and then entered by a person without any substantive knowledge of the data, this person could only enter a "missing data" code and refer the matter back to the person collecting the data. In the latter case, the correct value is more apt to never be entered. With this in mind, the following features of a system are most important:

a) The computer should display a familiar data entry form on the screen, and the operator should be able to simply fill in appropriate values in the correct blanks. An electronic cursor should automatically jump from field to field as values are entered or a special key is pressed indicating that a particular field is to be skipped. All instructions, such as which key is to be pressed to accept a value, reject a value, or skip to the next field should be clearly displayed on the screen.

b) The data entry program should be designed so that the operator need only type in <u>raw</u> data; all calculations using these data to yield values such as time-weighted averages should be done automatically. Similarly, the program should allow easy conversion from one type of unit to another; e.g., mg/M^3 to ppm. This could be done either on the standard screen or on a "worksheet" screen or subscreen that would be called up by pressing a special key on the terminal. Again, the use of all keys should be described on the screen and the final value inserted into the appropriate field automatically.

c) As data are entered, the program should check for invalid responses and do at least some preliminary checks for their accuracy. If a category code is expected, the machine must check to see that the value entered is a valid category for that field (e.g., if the codes run from 1 to 20, a value of 24 should be rejected and the user asked for a replacement; similarly it should check to see that social security numbers have nine digits, etc.). Further checks are possible, albeit requiring more computation time. For example, not all nine-digit social security numbers are typed correctly. The program could check to see not only that the number has nine digits but also that those nine digits match the identification number of an employee of the base for which data are being entered. If no match is found, the user should be notified that a problem exists, although the machine probably could not determine if the number was typed incorrectly or if the master personnel list was incomplete. This logic could be applied to other identification codes, such as workplace identifiers, chemical codes, or equipment codes. Values obtained from sampling can take on a virtually infinite range in principle, but some quick checks on plausibility might help catch simple transcription or typing errors. The program could, for example, check the new value against previous values. If the two values differed by a very large margin, the operator could be warned to double-check.

d) Ultimately, the system should allow data to be read directly from biological and industrial hygiene monitoring equipment (e.g., audiometer readings), so that results can be transferred directly rather than being transcribed by a human and then entered via keyboard. This should help improve the accuracy of the data in the system by eliminating opportunities for error.

Information Retrieval

The system should allow easy access to data bases stored on other computers, in and out of the Air Force, that contain information of use to the various types of professionals for whom the system is designed. Ideally, the user would use these different systems under the control of an occupational health computer that would allow queries and print responses in a single, user-oriented fashion. This has been suggested[1] as a way for inexperienced users to make use of several computer systems through a single, simple command language. Computers with a layer of "protective ware" could be used as communications interfaces that would translate simple instructions from the users into the probably more complex and certainly less standardized host operating system commands. They could also help interpret the often cryptic and jargon-ridden responses from the host computer in a way that would help the user to respond effectively. Several on-line data sources listed below would be useful in this connection:

a) A hazardous materials information system (HMIS), which would contain data on the names, chemical properties, and known hazards of materials used by the Air Force. This should be a source for threshold-level (TLV) information on materials of concern that would be more timely and easy use than printed sources.

b) The National Library of Medicine Toxicology Data Bank.

c) Results of sample analyses, film-badge reports, etc., sent to OEHL for processing.

d) An automated medical records system for employees should be accessible to the occupational health computer system.

e) The Air Force hearing conservation data base.

Data Analysis/Report Generation

It is very important that the system be capable of relating the various types of data to one another so that reasonable analyses can be performed. Specifically, links should be designed between:

a) an individual and his/her workplace

b) an individual and his/her health records (including personal radiation exposure data)

c) chemical identification/known hazards/handling guidelines and workplaces where the chemical is used

d) discrepancies from handling guidelines involving workplace location and chemical exposure

[1] E. B. James and D. Ireland. Microcomputers as protective interfaces in computing networks. Software Pract Experience 10:953-958 (1980).

e) sample identification and workplace

With this scheme, direct or indirect links can be made between any two pieces of information; the workplace identifier is often used as the intervening link. This scheme assumes that the definition of "workplaces" is done with reference to the similarities in processes, hazards, and controls encountered by a group of employees. If workplace identifications are done on a more arbitrary basis, the validity of some of the logically possible linkages will be questionable.

Once these fundamental requirements are met, the following data analysis capabilities are most important:

a) Users at the bases and the Air Force level should be able to make ad hoc queries of the data in a flexible manner. For example, a BEE or health care provider should be able to easily obtain a list of all workplaces in which a particular chemical is used, or to obtain the names of all employees at a particular workplace who received radiation exposures above a certain level.

b) Summary reports of hazards, accidents, and illness should be produced regularly, with percentages broken down by relevant characteristics of the workers, the workplaces, and geographic location.

c) Statistical analyses of associations between suspected hazards and morbidity-mortality data should be easily produced by relevant professionals. For example, health care providers at the base level should be able to perform preliminary statistical analyses on suspected hazards. If they determine that the evidence warrants more in-depth study, epidemiological specialists at some central location should be able to perform the appropriate analyses.

Physical Storage Requirements

Many of the analyses discussed in this report require some estimates of the physical volume of data to be stored. This depends, of course, on a number of decisions that have yet to be made, including the number of items on the SOHP forms to be computerized, the type of data management software used, and the types of reports that must be routinely generated. In order to secure reasonable values for this report, estimates have been made of the number of characters needed to store the information on the SOHP industrial hygiene and occupational medicine forms. In actual practice, a form would not be completely filled, and a reasonable storage scheme could take advantage of this fact. Furthermore, much numeric data and textual data that can be grouped into a relatively small number of categories can be stored more compactly than is assumed here. Thus, simply counting the number of characters that can be entered on the forms will tend to overestimate the amount of storage capability required. On the other hand, modern data storage schemes that allow flexible and rapid access to particular data values do not pack information into memory as tightly as do more simple-minded schemes; additional space is used by bucket storage/hashing schemes, pointer chains, secondary indices, etc. Since all of these factors cannot be estimated without knowing exactly which data and programs will be used, the estimates presented here are believed to be reasonable approximations.

From a count of the spaces on the forms to be stored in Tabs A-E of the new case files, approximately 16,000 characters (bytes[2]) would be needed to store the information on one workplace for a given year. Assuming[3] an average of 100 workplaces/base, 1.6 MB/year would be required to store industrial hygiene data for an average base.

By a similar counting of the occupational health forms, Tab F will require an additional 6000 bytes of storage space per person. According to estimates supplied to this study group, an average of 150 complete physical examinations are performed on each base in a given year, requiring an additional .9 MB annually. Most occupational health examinations, however, are to monitor hearing loss. Assuming that 300 bytes are required to store this subset of the physical exam data and that an average of 3000 are performed at an average base in a given year, another .9 MB/year would be required. Thus, 3.4 MB/year/base, or 17 MB over a 5-year period, would be needed to store the industrial hygiene and occupational health data for a single base. An additional 3-5 MB would be required to store data on the most common chemicals used on a base, so a reasonable estimate of the storage requirements of a typical base would be 20 MB over a 5-year period.

[2] Computer memory is measured in bytes; that is, the number of alphabetic characters, numeric digits, or machine instructions that can be stored. A kilobyte (KB) is 2^{10} (1024) bytes, a megabyte is 2^{20} (1,048,576) bytes, and a gigabyte (GB) is 2^{30} (1,073,741,824) bytes. Microcomputer main memories average 48-64 KB and usually have floppy disk drives with 100-300-KB capacities. Minicomputers tend to have 128 KB to 1 MB of main memory and 10-100 MB of disk storage. Mainframes tend to have 1-20 MB of main memory and up to several GB of disk capacity.

[3] The assumptions used here are based on conversations with personnel at USAF OEHL.

CHAPTER 2: SYSTEM HARDWARE/SOFTWARE CONFIGURATION

Centralized or Decentralized Processing?

This section outlines several architectures for the Air Force occupational health data processing system. These vary mainly in the degree of centralization of the system. Considered possibilities range from acquiring a large central computer that all bases will use via "dumb" terminals to a totally decentralized system in which the OEHL computer would simply be used as a data repository. While not all of the scenarios outlined merit in-depth consideration, the discussion should help illuminate the range of technologically feasible options and the implications of each for the actual problem at hand.

The following configurations are discussed in some detail:

a) Complete centralization--A central computer is used for all data entry and processing; bases interact with it via ordinary computer terminals and phone lines.

b) Partial centralization--The central computer is used to store and analyze data, but users interact via microcomputer "smart" terminals that support data entry, data transmission, and some simple analyses.

c) Two-level distributed data base--There is a network of minicomputers; data are stored and analyzed locally, but the central computer can automatically request specific data for Air-Force-wide analyses.

d) Decentralization--Virtually all processing is done at base-level computers; there is no direct tie-in to a central computer. Data requests must be processed by humans at base level.

In general, centralization of the data base favors the managers and epidemiologists at the Air Force level who need rapid access to data from any given base or person or for an Air-Force-wide study. Decentralization tends to favor the base-level personnel, who only need to use data relevant to one base and whose programs would be hampered by communication delays and the necessity of finding data for one base amidst those for more than 100 others.

Complete Centralization

Under this scheme, a large central computer, presumably at Brooks AFB or at the San Antonio Data Service Center (SADSC), would handle all data processing from data entry through report generation. Users at other bases would communicate via dumb terminals connected through a telephone system and/or a digital communications network such as Telenet or Tymnet. The central computer would need a time-sharing operating system capable of handling at least 30 users simultaneously. This assumes that 117 bases use the system an average of 3 hours per day,[4] allowing for time-zone differences and assuming that connect hours can be scheduled efficiently. The data storage requirements of such a system would be considerable: approximately 20 x 117 = 2340 MB/5-year period, given the assumptions just described.

This approach may seem cumbersome at first glance, but it does have advantages. Indeed, this basic configuration is currently used by at least one private corporation roughly equivalent to the Air Force in number of employees and geographic diversity of operations. The most obvious advantage is that communications are straightforward. The computing industry has extensive experience in designing equipment and software to interface large computers with remote terminals, even over long distances. Overseas communications would be more problematic: costs would be greater and there are incompatibilities between European phone systems and American communications equipment. An advantage of the centralized system is that once the data are put into files in the central computer, they do not have to be moved. Personnel movements pose no problems, since all data would be centrally located. Furthermore, many programs for analyzing the data can be easily obtained for the large mainframe computers needed to handle the data storage task.

A second advantage is that all support personnel (probably two people to operate and maintain the computer) could be centrally located, while more decentralized data systems often require replication of technical personnel at each location. Thus, skilled personnel costs are likely to be lower for a centralized system.[5,6]

Third, there are economies of scale in data storage, even though the cost advantage of large computer processing has declined in recent years.

---

[4] In conversations at OEHL, the study group was given the estimate of 3 hours use per day.

[5] Grayce M. Booth. The Distributed System Environment: Some Practical Approaches. New York: McGraw-Hill (1981).

[6] R. A. Davenport. Distributed or centralized data base. Computer J 21:1 (Feb 1978).

The principal disadvantage of such a system is communications cost. Telecommunications costs have tended to rise with the rate of inflation even as computer hardware costs have decreased. At United States long-distance rates, the costs of communicating between the bases and the central computer would be prohibitive. Telenet or Tymnet (which offer a higher quality connection) costs are more reasonable, but still high: Assuming 3 hours/day/base and 260 working days per year, the 117 Air Force bases would spend more than $775,000 per year on occupational health data communications alone (assuming a rate of $8.50/hour,[1] and this would certainly be higher for overseas bases).

Other disadvantages relate more to convenience than economy. One is speed of communications. Slower turnaround time (i.e., the time between the user typing the last character and the first character of the machine's response) is more or less inevitable when the signals themselves must cross the country, passing through many levels of switching mechanisms. Similarly, with all users connecting to only one computer, all data processing would come to a standstill when it malfunctions.

To implement a simple, automatic method of accessing computerized records stored on other computers (e.g., the NLM Toxicology Data Bank) would be difficult under this scheme. The BEEs and health care providers would dial directly into those computers in the conventional manner rather than having the details handled by the central computer as discussed under Ideal Specifications in Chapter 1.

Partial Centralization

This scheme would closely resemble the centralized system, but the data entry would be accomplished via microcomputer-based smart terminals at the bases rather than dumb terminals connected directly to the host mainframe. In other words, the microcomputers would act as "front-end processors" for a mainframe data base management system (DBMS). Variations on this basic scheme could assign more data analysis capability to the local microcomputer terminals, but all data would ultimately be stored on the central computer. Thus, most of the specifications for the central computer would be the same as for the completely centralized option, except that the number of users who would require simultaneous access to the host computer would be reduced.

[1] Until recently, Telenet charges averaged about $5.00/hour and Tymnet charges were about $8.00/hour. Both are about to increase considerably; for example, Telenet charges to the University of Michigan computing system have recently risen more than 50%, to $8.50/hour.

The local microcomputer could perform several tasks, the exact number depending on the amount of storage capacity available to it. The most important task of this microcomputer would be to handle data entry, data editing, and communication of data to the host mainframe. For example, a system could be developed to allow a user to update industrial hygiene data for several workplaces by activating the microcomputer and telling it which workplaces need modification; it would then establish a link to the host computer and retrieve the current copy of those data (if any) into its own memory. The microcomputer would then break the link with the host machine and assist the user to make changes to the data, add new data, delete old records, etc. This program could also check new data for logical errors and warn the user when rapid changes or unacceptable levels of some value are entered, as well as assist in routine calculations and conversions. If a printer were available with the microcomputer, printed copies of the data could be made for final checking and manual storage. When all changes had been made, the microcomputer would reestablish a link to the host computer, transmit the updated information, and then sign off.

The microcomputer could also "know" the phone numbers and login/query procedures for other centralized data bases of interest to occupational medicine and industrial hygiene personnel; thus it could operate the protective interface software.

If the local *microcomputers* were given a fair amount of storage capability of their own (e.g., a 10-MB hard disk), at least the industrial hygiene data could be stored locally and only updates would be regularly transmitted to the host. This would save the cost of retrieving the original copy of some data from the host. If such storage capacity were available locally, most of the BEE's work with the data could be done off-line from the host. The data base could be queried and reports generated without making contact with the central computer. Since a base has many more people than workplaces, storing occupational medicine data on the microcomputer may not be feasible; thus these data would be stored centrally but edited locally.

The principal advantage of such a system over the completely centralized system would be that data transmission costs would be cut considerably. Rather than an average of 3 hours per day per base of connect time, 1 hour is a better estimate.[a] This would save not only on actual data transmission costs, but also on hardware for the central computer since it would have to handle a smaller number of users at a given time. Also, giving the local terminal some intelligence allows the microcomputer to connect to and deal with the host computer without much human intervention; thus many transmissions could be queued until a port on the host was available or until off-hours when transmission costs are lower.

Such a system would also be more available to the average user. Actual sessions with the host would be of short duration (waiting time would tend to be short), and an interruption in the service of the host would still allow work to be done on the base microcomputers.

Another advantage is that considerable off-the-shelf hardware and software is available for at least the more popular microcomputers. Hardware modifications and programs for telecommunications are available for many models, as are programs for data base management, text processing, and other functions which the USAF Occupational Health Program plans to have available. Nevertheless, the sheer number of some kinds of microcomputer systems has led to the widespread availability of powerful language compilers and a more widespread familiarity with the systems among programmers. This can be expected to facilitate software development.

Many microcomputers have more sophisticated graphic display capabilities than would be feasible with a fully centralized system. These could be used to allow more powerful and easily interpretable summary displays of various data, such as time plots, bar charts, etc., and would be useful for clearly displaying workplace diagrams.

The major disadvantage of this partial centralization scheme is that the cost of the local microcomputer--its terminal, printer, disk equipment, communications hardware, and the necessary software--would be on the order of $10,000 rather than $1500-$2000 or so for a good but basically dumb terminal. Such a system would impose a fairly high initial

[a] Assuming that a session consisted of retrieving four 16-KB workplace records, editing them, and returning them to the central computer, 8-9 minutes/record would be needed to transmit them each way at 30 characters/second. The efficiency of this transmission could be further enhanced if only the parts of a record actually modified are returned to the central computer. Thus, 36 minutes in one direction and 24 minutes in the other is a reasonable estimate of daily connect time with the central computer.

hardware cost and would not totally eliminate communications charges.

Also, maintenance of the components of the system would be more difficult than under a centralized/dumb-terminal configuration. Microcomputer systems (especially their peripherals) are simply not as troublefree as CRT terminals.

Distributed Data Base

This system configuration would consist of a two-level hierarchy of minicomputers. All raw data would be stored at the base level, but the network data management software would allow a central user to view the data as being "logically" integrated. Thus, all data would be kept on-line at individual bases and would be more or less immediately available to users at other bases via the network software. Each computer in the network would be fairly independent of the others, so a failure of one would not affect the others (except, of course, that data could not be retrieved over the network from a subsystem that was down). The network hardware and software would be used so that specific data items from a particular base could be quickly and automatically forwarded to a central computer as they were needed. In other words, the processes of finding and transmitting data from computer to computer would be invisible to the central user, who could view the data as if they were all together. The software would handle the details involved in physically carrying out instructions. This process would be facilitated by a systemwide "data dictionary" on the central computer that would direct queries to the appropriate subsystem without having to search through the files of 117 machines for a particular item. For example, a mechanism could be provided to transmit the complete record on a person up and down the hierarchy to follow a person as he or she is transferred within the Air Force.

Such a system would require minicomputers at each base, with communications interface hardware on each computer. The central computer would typically be a "super" minicomputer or mainframe with enough capability for analysis but not storage of the raw data. The most typical requirements would be about 1 MB of main memory and 100 MB of disk. The central computer would also require network processing equipment.

As for software, both the base and central computers would require a full-scale DBMS. The base computers would need interactive input and reporting programs; and if the DBMS did not have an on-line query facility, an interactive data retrieval program would be required. The central computer would need facilities to receive abstracted data automatically from the bases (or generate particular queries over the network) and would require a fairly powerful DBMS with an interface to a statistical package.

Such a configuration has distinct advantages. First, this type of distributed system would keep communications costs at the minimum level that could be achieved with any system of direct communication between the central and base level. This would be possible because most data used on a day-to-day basis would be stored on a computer at the base it represents, so routine transactions would be handled locally. When data were required from another computer, they would be sent at a high transmission rate, minimizing connection time. Also, this configuration would be "up" a higher percentage of the time from the perspective of the average user, since problems with the central computer would not hinder most users at the base level and problems with a base machine would not affect users at other bases (except on the relatively rare occasions that data from one base are needed at another).

On the other hand, this configuration has a number of potential disadvantages. First, automatic distributed data processing is still very much a state-of-the-art design problem.[9] Several vendors have developed hardware and software to communicate among their own machines in the manner described here, but whether or not the bugs have been worked out is not clear. Communications problems are more likely if the central computer is supplied by a different vendor than the base-level computers. This situation might arise if, for reasons of economy, an attempt is made to incorporate existing hardware into the new system. Also, personnel costs would tend to be higher with a distributed system, since a local technical staff would be required to maintain a minicomputer system of the appropriate size (probably one person per base system). Such costs would, however, be justified across the various organizations on each base using the minicomputer. In addition to the hardware costs, a strong possibility exists that software licensing requirements would necessitate that programs purchased just for the occupational health system would cost considerably more if used on 117 computers than if used on just one.

Decentralization

A completely decentralized system would be very similar to the distributed system described, except that base computers would have no direct links with the central computer. Abstracts of base-level data would have to be forwarded regularly to the central computer via magnetic tape or

[9] Grayce Booth. The Distributed System Environment, p. 35.

floppy disk and then would be manually incorporated into the data base for managerial and epidemiological uses. Special queries could be handled, but would have to be processed manually at both ends.

The only advantage of such a system would be cost. It could make considerable use of existing and planned computer equipment (as might the distributed system), but would not impose state-of-the-art technological requirements on the communication software. Such a system might be upgraded to a distributed system when finances and technological advances permit. Also, the minicomputers on which such a system could run generally do not require a local technical staff to operate them.

The chief disadvantage would be a considerable lack of flexibility. Instructions for abstracting and forwarding data would need to be developed in advance, and special requests for the forwarding of needed but unanticipated data would be cumbersome to provide. Furthermore, this system has no way to help the users connect to and query other data bases; such queries would have to be carried out manually with the dumb terminals used with the base minicomputers.

Software Available

Complete Occupational Health Systems

These are large programs or sets of programs that are intended to handle all data processing associated with an occupational health surveillance system, from data entry through data analysis and report generation.

Diamond-Shamrock's COHESS--To handle its own occupational health data management needs, Diamond-Shamrock designed this program, which has been sold to several other organizations (including AFLC). It is quite flexible: It can handle three different types of data (on people, places, and things), and the user can specify what exactly is to be monitored in each. Under "People," data such as basic personnel records, health evaluation results, and health incidents (e.g., clinic visits, absence reports, accident records, and morbidity-mortality data) can be stored. The "Places" data include an index of workplaces by location and the results of various industrial hygiene monitoring efforts. "Things" data consist mainly of data on potentially hazardous materials used in the organization.

COHESS is set up to produce two types of reports. First, it handles the scheduling of physical examinations. Second, it allows various types of ad-hoc queries of any of the data. For example, a user can ask for the names of all people meeting some demographic criterion who have been exposed to a certain chemical and have values of some clinical test above a particular value.

COHESS has rather restrictive hardware requirements. The master data base is stored on a large IBM-type computer using the IMS data base system. The user interacts with a program on a Datapoint minicomputer that processes queries and passes them on to the mainframe, then stores and formats the response. The minicomputer portion of the program is written in a proprietary language called Datashare, so it could not be easily transported to another type of machine.

EVALUATION: COHESS can handle most of the high-priority tasks outlined in Ideal Specifications, Chapter 1. It is a flexible program, and Air Force data coding schemes and forms could be used with the system. Much of the data (the so-called dumb data) is entered via CRT terminals displaying predefined forms. It is not, however, very "user-friendly": some of the data entry (the smart data) must be done by specialists. No mechanism is available to help the user do the routine calculations and conversions often encountered, and its ability to "intelligently" check incoming data for errors is limited. The fact that most of the system is written in a proprietary language is a very severe limitation. In essence, the Air Force would have to buy Datapoint computers for each base or go to considerable effort to translate the program into some other language. This effort would surely be better spent in developing a system tailored to the needs of the Air Force.

InSci's "Expanded OSHA-Health/Safety System"-- Information Science, Inc. markets an occupational health monitoring system designed to be used in conjunction with other packages it provides for data base management and data analysis. This system is more rigidly structured than COHESS. It handles hree types of data: accident/illness records, industrial hygiene monitoring data, and health records (including physical examinations, clinic visits, environmental exposure, and work history). The outputs include the OSHA 200 form (log and summary of occupational injuries and illnesses), employee accident/illness histories, physical examination profiles, average exposure by substance, and hearing loss reports. There is also a "General Retrieval System" that can be used to query the data base. It does searches, selects subsets, performs calculations, and displays information in response to commands in a lanouage designed to be used by nonprogrammers. This query language, however, appears to be more clumsy than those used in many other commercial DBMSs. Epidemiological studies can be done with the same data base using yet another package.

The system requires an IBM 360/370 computer with an OS, DOS, or VS operating system, 250 KB of main memory, and the equivalent of five tape drives and two disks.

EVALUATION: This is a batch-oriented, sequential storage system that uses the software technology of the late 1960's. It probably meets the needs of many medium-sized businesses, but does not meet the basic requirements for an Air-Force-wide system because it is not interactive, is not very user oriented, and has no provision for intelligent error-checking of data as they are entered.

S. C. Johnson's "Occupational Health Information System"--This program (OHIS) is currently being marketed by Environment Systems Specialists under license from Johnson. It handles three basic types of data: personnel (including demographic information, work histories, work locations), medical (including history, examinations, and morbidity/mortality), and industrial hygiene/toxicology (including survey records, toxic effects information, and material handling guides). The exact contents of all three can be tailored to the needs of a particular installation without extensive reprogramming.

The system has reasonable data input capabilities. Iata can be entered onto standardized forms displayed on the computer's screen, the program handles the routine calculations used in industrial hygiene, and the system does extensive error checking as the data are entered. For example, the program attempts to look up any identification codes as they are entered for a person or workplace. If no match is found, the operator is warned and prompted for corrections. The system at Johnson currently gets some data directly from microprocessor-controlled laboratory instruments (for pulmonary function and hearing tests), thus reducing data input errors even further. All laboratory data are checked to see if they have been requisitioned, values are compared to normal ranges and permissible limits for quality control purposes, and a standard laboratory test report is generated.

The system can produce many types of output, including periodic standard reports, life-stress analyses, physical examination schedules, and ad-hoc queries. Most importantly, the various data can be combined to yield correlations between, for example, industrial hygiene exposure data and health records.

It is written in the Digital Standard MUMPS language, which is a variation of the ANSI standard language widely used for medical applications. In principle, this system could be run on any system with a MUMPS interpreter; in practice, the program is tailored to Digital Equipment Corporation (DEC) equipment with at least 128 KB of memory and 10 MB of disk storage. Two basic configurations for MUMPS systems exist on DEC equipment. In the one used by Johnson, MUMPS is essentially the operating system of a fairly small minicomputer. In the other, MUMPS runs as a subtask on a VAX super minicomputer; under this configuration, MUMPS programs can call subroutines written in FORTRAN or other languages more suitable for graphics or numerical work.

EVALUATION: The OHIS handles the high-priority tasks described and is user-friendly and easily modifiable. It already has many capabilities that are close to the ideal data-entry specifications (chapter 1). While a more detailed evaluation of the actual program will be necessary, apparently it can be modified to suit the needs of the Air Force at somewhat less cost than developing a new system. One major drawback is apt to be the relative inefficiency of MUMPS for programs run on a routine basis. This is probably outweighed by the advantages of the language in terms of easily modifying the OHIS program to meet Air Force needs and the ease of writing special-purpose programs using the USAF Occupational Health data base. Further evaluation of the actual code for this system is warranted to determine its actual suitability for Air Force needs.

Data Base Management Systems (DBMS)

The core of the occupational health software system is a DBMS: A software package that allows data to be stored in a secure format, makes it easy for authorized users to retrieve and modify particular pieces of information, and allows other programs to readily access its data for multiple applications without detailed knowledge of its internal workings. In other words, a DBMS provides a flexible, general-purpose filing system that should considerably reduce the tedium involved in storing, updating, and retrieving computerized information.

Due to the number of packages available and the uncertainties about the hardware on which the system is to be run, this study group cannot make firm recommendations on exactly which DBMS is most appropriate for the needs of the Air Force; but some general issues are worth noting. Data base systems can be roughly categorized into two groups: those best for "static" applications and those more suited for "dynamic" applications.[10] The former include most routine business applications in which the same program is run on a large data base on a regular basis; e.g., producing a payroll. Dynamic applications are more ad hoc and less predictable. In this case, data are merged and queried in ways that may not have been foreseen by the designers of the data processing system. In terms of the issue at hand, routine analyses such as breakdowns of illness statistics by base, workplace, or job classification would be essentially

[10] Michael M. Gorman. Choosing the right DBMS for your application. Computerworld 15(26):10-12 (June 29, 1981).

static; while less structured queries of the data base by field personnel would tend to be dynamic.

The DBMS packages most suited for static applications tend to be based on the hierarchical or network data model, in which data are retrieved by chains of pointers that must be specified when the data base is designed.[11] Systems of this sort include IBM's IMS, Cincom System's Total, and Intel's System 2000. Those best suited for dynamic applications are usually organized according to the relational data model and/or use inverted-list physical organizations to facilitate fast and flexible retrieval. Representative systems include Software AG's Adabas, Mathematica's RAMIS, and the forthcoming IBM System R. While DBMSs of each type are available on systems of all sizes, standards to ensure compatibility across machines are few. Thus even if, under the partially centralized configuration, relational DBMSs were acquired for both the base-level microcomputers and the central mainframe, special communication software would have to be written to translate records from one into another. The vendors of distributed network equipment and software maintain compatibility between the data schemes on the large and small computers in the network, but these usually employ hierarchical DBMS storage schemes.

For the base-level data bases that will presumably be stored on mini- or micro-computers, the applications probably will be primarily dynamic; hence, a data base package organized along relational or inverted-list lines would be most appropriate. Such packages are becoming popular on computers of all sizes, so availability should be no problem. Likewise, if the S. C. Johnson OHIS package is purchased, the choice of the data base scheme has already been made by the designers of that program; the program was designed to make best use of the MUMPS DBMS. If a centralized configuration is chosen, however, it will be necessary to specify in advance the retrieval paths that will be used in the routine analyses. Efficiencies in the use of computer time can be achieved by using a hierarchical or network DBMS. Ad hoc queries may be performed and subsets generated that were not anticipated. These queries can indeed be done even with a static system, but would be less efficient than under a more dynamically oriented DBMS package.

In many situations, both routine and ad hoc queries are commonly made. In such cases, maintaining two more-or-less identical data bases in parallel may be more efficient. For example, the oil company examined under Task 2 of this contract maintains occupational health data under IBM'S IMS

[11] James Martin. Computer Data Base Organization. Englewood Cliffs, N.J.: Prentice-Hall (1977).

for long-term storage and routine report generation. The most current of these data are periodically converted into the more dynamic RAMIS data base format for use by industrial hygienists calling in from the field who make more restricted but less predictable queries.

Microcomputer Software

A popular new family of programs, the best-known example being VisiCorp's VISICALC, are useful for flexible data entry, display, and updating. Such programs are now available on virtually all microcomputer systems and should be useful as a relatively cheap but effective system for entering industrial hygiene data. Unlike conventional DBMSs for microcomputers, these programs are very visually oriented, which makes the data easy to find and edit. Furthermore, calculations can be easily programmed, and simple totals and subtotals can be set up quickly. VISICALC itself has spawned a large number of utility programs that perform more sophisticated data base management functions, computer-to-computer communication of data bases, and report/display generation.

CHAPTER 3: SYSTEMS DEVELOPMENT

Computer Language Considerations

Even if "canned" programs are used extensively, some computer programs will have to be written to manage the USAF Occupational Health Program. Thus, the perennial question of which computer language is most appropriate deserves attention. The most common language in business and the Government is COBOL, which is highly standardized (i.e, a standard COBOL program written on one machine will, in principle, run on another without modification). However, this language has some well-known deficiencies:

a) It is rather verbose and clumsy to work with, meaning that programmers operate at a lower productivity rate than if they were using other languages. It is not naturally block-structured, so writing programs using the modern discipline of "structured programming" is somewhat awkward.

b) It is not very efficient for numerical applications; that is, for the types of more sophisticated mathematical techniques used by statistical analysis and graphical display programs.

Since the limitations of COBOL are well known, several attempts have been made to devise new languages that preserve its strengths but correct its weaknesses. The Department of Defense (DOD) has recently commissioned the development of a new language, known as Ada, that is expected to replace COBOL as the DOD standard. Release of the definition of Ada in late 1980 has led to much activity in the computing industry to prepare Ada compilers, so it is possible that the language can be used to develop the USAF Occupational Health Program software.

The problems of COBOL are addressed by Ada: It is powerful but terse, so programmer productivity should be increased, and it is designed to be efficient for both numerical computation and business data processing.[12] But caution is in order here: this new, untried, and somewhat complex

[12] There is a conventional distinction between "scientific" languages (requiring binary arithmetic, floating-point arithmetic, double-precision variables, complex variables, and a good built-in library of mathematical functions); and business languages (requiring good file-handling capabilities, decimal arithmetic, good character data handling features, and easy output formatting). Business applications now often require scientific features, and vice-versa, so modern languages at least attempt to provide both sets of features.

language may be difficult to implement effectively. DOD opted for a powerful language at the expense of simplicity, and it is possible that Ada will suffer the fate of such languages as ALGOL68 and PL/I, which attempted to do too much and essentially failed.[13]

Ada's predecessors are still in use and have been developed to provide useful alternatives to COBOL for general-purpose data processing. PL/I Subset G incorporates many of the strengths of full PL/I but avoids some of the more problematic complications. Pascal is another modern language that is very useful for easily implementing logical, well-structured programs, but the standard version is not very useful for business data processing. Several non-standard versions of Pascal avoid this problem, especially the version developed at the University of California at San Diego (the UCSD P-System), which is available for most microcomputer systems.

Ada is based largely on Pascal, although it has many features not found in the earlier language. Translating a program from Pascal to Ada is, in principle, a fairly straightforward matter of changing syntax.[14] If Ada is not available on one of the computers to be used for the occupational health system, an interim solution could be to develop the system in Pascal, then upgrade it when an Ada compiler becomes available. Since the specifications for Ada are already public, this would simply require some discipline on the part of the programmers to ensure that their code is compatible with Ada. In this way, the system could be developed in a modern, structured language rather than COBOL, yet would ultimately enjoy the advantages of standardization. Moreover, this would avoid using the parts of Ada that are likely to be problematic if Hoare's gloomy assessment is correct.

Other, more piecemeal solutions have been devised to circumvent the problems of COBOL. One is the use of "code generators" which help the programmer work more efficiently by automating much of the tedium of coding COBOL programs. At least one company investigated by this study group used this technique to implement a COBOL-based occupational health data system in relatively few person-years. Data base management systems are useful in this regard both because they simplify the process of programming data retrieval and because they can provide a bridge between

[13] C. A. R. Hoare. The emperor's old clothes (ACM Turing Award lecture). Commun ACM 24(2):75-83 (1981).

[14] Paul F. Albrecht et al. Source-to-source translation: Ada to Pascal and Pascal to Ada. ACM SIGPLAN Notices 15(11):183-193 (1981).

COBOL data entry and manipulation programs and data analysis programs written in other, more suitable languages. The petrochemical company investigated under Task 2 of this project used this technique to link together its COBOL-based data management programs and its prepackaged data analysis programs. The master files are maintained with COBOL programs that use IBM's IMS data base management system. These data are periodically copied to a RAMIS data base, which can be accessed by the analysis software; e.g., SAS. In principle, the master data files could be maintained by a DBMS that can also be accessed by programs written in another language, thus eliminating the conversion step.

MUMPS is a standardized language widely used for medical applications. Unlike the other languages discussed here, MUMPS is implemented with an interpreter rather than a compiler, meaning that statements are only translated to machine code when a particular line of MUMPS code is executed, whereas a compiler translates an entire program at once. While compiled programs run faster than interpreted programs, programs in interpreted languages are easier to modify and debug. Another powerful feature of MUMPS is that it has considerable data base management capability built into the language itself, thus making the separate purchase of a DBMS unnecessary. On the other hand, MUMPS requires special effort in writing well-structured programs. As with COBOL, reasonably clear, maintainable code can be written, but this requires more use of advanced features of the language and careful management of the programmers whereas Pascal or Ada make it natural to write clear, easily understandable programs.

Software Development Strategies

An entire field known as software engineering has recently developed. Several important principles have emerged that can be summarized briefly as follows:

First, system design should proceed in a "top-down" manner. That is, the modules that control interaction with the user and the production of useful output should be specified first in an abstract way, then gradually fleshed out by a process of stepwise refinement. It is very important for the actual users to be represented in the early phases of the design process, so that they clearly specify what they require from the system. The lower-level design issues can be left to specialists.[15]

[15] D. J. Mishelevich and D. VanSlyke. Application development system: The software architecture of the IBM Health Care Support/DL/I-patient care system. IBM Systems J 19:4 (1980).

Second, the system should be implemented and tested incrementally rather than all at once. That is, no attempt should be made to design and code the entire system in one step, then throw its subcomponents together and expect them to work. No matter how meticulous the design process, this is sure to complicate the debugging process.[16]

Another principle is thɑc one must resist the temptation to use manpower to buy time in software-development projects. F. P. Brooks is the best-known exponent of the opinion that "adding manpower to an already late software project makes it later." This is because the time involved in designing and implementing software is largely taken up by basically creative and integrative processes for which conceptual unity is important; adding additional personnel makes the task of coordination more difficult and tends to decrease the conceptual unity of a program.[17]

Similarly, one should appreciate the relationship between the time it will take to complete coding and debugging a software project and the amount of time spent in planning. Many believe that detailed planning at the beginning will cut debugging time far more than it will delay the beginning of actual coding.

Privacy Considerations

Considerable attention has been given in the last decade to the protection of the rights to privacy of individuals about whom computerized records are maintained. The main problem stems from the fact that the occupational medicine records contain information that is necessary for effective monitoring and yet may be of a sensitive nature.

The word "privacy" is not completely clear; this discussion makes three basic assumptions about privacy rights:

a) Individuals are assumed to have a right to participate in determining how information about them is to be used by a data-collecting organization and under what circumstances such information will be transmitted to others;

b) Individuals should be assured of openness, forthrightness, and fairness in record-keeping;

[16] Edward Yourdon and Larry L. Cσnstantine. Structured Design, p. 501. New York: Yourdon (1975).

[17] Frederick P. Brooks, Jr. The Mythical Man-Month: Essays in Software Engineering, p. 21. Reading, Mass.: Addison-Wesley (1975).

c) Individuals should be protected against unwelcome, improper, or excessive data collection.[18]

The Federal Privacy Act of 1974 gave legal force to a growing public sentiment for privacy protection. An excerpt from section 2(b) of the Privacy Act gives a good sense of what it attempted to accomplish and summarizes its provisions:

"The purpose of this Act is to provide certain safeguards for an individual against an invasion of privacy by requiring Federal agencies, except as otherwise provided by law, to--

(1) permit an individual to determine what records pertaining to him are collected, maintained, used, or disseminated by such agencies;

(2) permit an individual to prevent records pertaining to him obtained by such agencies for a particular purpose from being used or made available for another purpose without his consent;

(3) permit an individual to gain access to information pertaining to him in Federal agency records, to have a copy made of all or any portion thereof, and to correct or amend such records;

(4) collect, maintain, use, or disseminate any record of identifiable personal information in a manner that assures that such action is for a necessary and lawful purpose, that the information is current and accurate for its intended use, and that safeguards are provided to prevent misuse of that information;

(5) permit exemptions from the requirements with respect to records provided in this Act only in those cases where there is an important public policy need for such exemption as has been determined by specified statutory authority; and

(6) will be subject to civil suit for any damages which occur as a result of willful or intentional action which violates any individual's rights under this Act."

Thus, the Privacy Act does not in any way preclude the Air Force from establishing a computerized data bank of personal health records. The Air Force, however, must ensure

[18] Willis Ware. Privacy and information technology: The years ahead. *In* Lance J. Hoffman (ed.) Computers and Privacy in the Next Decade. New York: Academic Press (1980).

the accuracy of the data, prohibit unauthorized dissemination of personal data, and give individuals the right to inspect their own records. The Act does constrain the Government's right to use occupational health data for other purposes, e.g., to obtain information on an individual's personal habits during the course of a security clearance investigation. However, statistical research is expressly permitted, so long as reasonable precautions are taken to protect the identity of the subjects.

Looking beyond the legal requirements of today's privacy laws, several authors have made recommendations for developers of computerized information systems that incorporate likely future directions in privacy protection law and practice.[19, 20, 21, 22] The gist of these recommendations can be synthesized into eight guidelines for system developers:

a) Prepare a privacy impact statement, which would be communicated to the population of individuals whose records would be automated. A study of computerized health records found that most public relations and legal problems related to privacy considerations stemmed from inadequate consultation in advance with groups representing citizen rights and a lack of proceedings open to the general public to explain and justify the data collection system. Such a statement and/or public forum would discuss (1) controls on the operating practices of the system, (2) access rights of the data subjects, (3) usage control by the data subjects, and (4) effects of privacy regulations on the operation of the system.

b) Construct a comprehensive privacy plan right from the beginning. This would make clear how privacy controls are to be integrated into the design of the system: Keeping this consideration in mind all along is cheaper than grafting controls on later and should ensure that no design decisions are taken that are inconsistent with privacy objectives.

[19] Alan F. Westin. Computers, health records, and citizen rights. NBS monograph 157, USGPO:C13.44:157 (1976).

[20] Gordon C. Everest. Non-uniform privacy laws: Implications and attempts at uniformity. In Lance J. Hoffman (ed.), Computers and Privacy in the Next Decade. New York: Academic Press (1980).

[21] Brad Schultz. Five Theses in Privacy. Computerworld 15(16):4 (20 April, 1981).

[22] Robert C. Goldstein and Richard L. Nolan. Personal privacy versus the corporate computer. Harvard Bus Rev, Mar-Apr 1975.

c) Inform individuals how the data they supply will be used. Informing subjects will demonstrate an organization's awareness and concern for the privacy of the data subjects and may also be of significant help later in obtaining data and the authorization to use it.

d) Provide some mechanism for individuals to have access to the information stored about them. Almost all privacy legislation gives subjects the right to inspect and challenge the accuracy of their own records. This creates some problems when one is dealing with medical information since much of it is unintelligible to the layman, but most authorities suggest that some mechanism be worked out to facilitate this.

e) Take steps to see that personal records are accurate, timely, and complete. This reflects a very widespread concern in privacy legislation that decisions which affect individuals must not be made on the basis of unreliable, inaccurate, or outdated information.

f) Provide a mechanism for limiting access to the data to only authorized users.

g) Conduct orientation and training programs for employees who handle personal information. Such programs should provide employees a respect for the privacy of the individuals about whom records are kept and familiarize them with the specific policies and procedures to be followed to ensure the confidentiality of those records.

h) Keep track of disclosures of personal data, recording whether they are of a routine or nonroutine nature. This would apply only to disclosures of data that can be linked back to a specific individual; statistical data would not require such surveillance.

CHAPTER 4: CONCLUSIONS AND RECOMMENDATIONS

General Recommendations

The principal assignment of this study group was to investigate occupational health data management systems developed by other organizations that may be appropriate for the needs of the U.S. Air Force. In the process of developing this report, it became apparent that the OHIS, developed by the S. C. Johnson and Son Company, could be of real value in meeting the USAF requirements.

This study group concludes that this software system could, with considerable modification, be appropriate to Air Force needs. The following estimates concerning the cost of these modifications and the cost of alternative systems are based on the best information available to the study group at this time.

On some issues there is little room for disagreement, and unconditional recommendations are presented first. The answers to other questions, however, depend on the system configuration that is chosen. Recommendations are made separately for each configuration proposed.

Interactive vs Batch Processing

Interactive processing is a virtual necessity if the system is to be useful to the BEEs and medical professionals in the field rather than be just a data collection device for central managers and researchers. Thus, base-level data used by local personnel should be on-line to the greatest extent possible. Having data for all bases available online to central users would be desirable (but not necessary). If this is not feasible, batch transmission of data to a central computer by tape or floppy disk would be acceptable.

Disk vs Tape Storage

Disk capability should be sufficient for data management and analysis software to use modern techniques based on random access to files. Some delays may occur while specific data needed by a program are retrieved from magnetic tape. Thus, storage hierarchies that use both magnetic tape and disk are acceptable, so long as the retrieval from tape to disk is transparent to the unsophisticated user.

Turnkey vs Programmers-only System

In a closely related matter, the basic data entry/ retrieval system at the bases should be oriented toward the end user (i.e., BEEs and health care professionals) rather than computer specialists. The system should be activated by simply turning on a terminal and issuing a small number of straightforward, easily remembered commands to the operating system. From that point on, instructions to the user should be in clear language, the required responses should be brief so as to not unduly tax the typing ability of the average person, and the output generated by the system should be clearly labeled and easy to understand. These three requirements are not difficult to meet; indeed, they simply define what is becoming widely accepted as the standard for user-oriented applications software in the 1980's.

User Access vs Confidentiality

In brief, the data should be accessible to any users with a plausible reason to have them, but the identity of the individuals on whom the data are collected should remain confidential to all but those with a real need to know such identity. Employees should be informed as to what steps are taken to ensure the confidentiality of the information they supply and how the data are used. A procedure should be devised so that individuals can inspect the records kept on them; mechanisms should be developed to limit the access of unauthorized persons to personal data; and those who operate the computer system should be trained to respect the privacy of the individuals about whom information is kept.

Single System vs Separate Military and Civilian Systems

Separating the two systems might have some short-run appeal. However, there is no fundamental reason to separate them since military and civilian personnel often work together at a particular site, are exposed to the same environmental hazards, and the BEEs and medical personnel must deal with both groups of workers at the same time. Separating the systems would make the process of entering and retrieving data more cumbersome to the end users. While this study group recognizes the practical difficulties involved in bringing civilian workers into the system, efforts should be made to do so.

In-house vs Contract Programmers

The study group cannot recommend simply contracting out the software development. The four corporations investigated had little luck with such arrangements; indeed, the software contractor who develops a system without a thorough knowledge of just what the customer wants has become a notorious stereotype in the computer industry. Effective software engineering must be done through close collaboration between end users and technical specialists. It is recommended that the Air Force retain the responsibility of designing whatever software packages are written for the

USAF Occupational Health Program. Outside consultants may be used in collaboration with Air Force personnel if more technical expertise is required on specific questions. Contractors can and probably should be used to actually code the programs from the detailed design directives. But here too, the contractors should be kept on a fairly short leash. An outside vendor might be entrusted with one or more modules of a large system, but this could be done on the basis of fairly short-term contracts with renewal contingent upon successful completion of the original task.

Configuration-Dependent Recommendations

The main issue, i.e., the one on which several others hinge, is that of centralization vs decentralization. Recommendations are made separately for three of the basic configurations discussed, and 5-year costs of storing and retrieving the base-level data are estimated.[23] Information used in making the estimates was obtained primarily from computer industry trade publications and from representatives of manufacturers. These estimates should be used only for comparing the relative costs of the various configurations; the rate of inflation and likely technological changes in computer technology have not been taken into account.

Partially Centralized System

In this plan, data would be entered through microcomputer smart terminals that handle data entry, checking, and some simple displays and then forward all data to a central computer for long-term storage and processing.

Canned Programs vs Custom Software--The study group cannot recommend any of the prepackaged occupational health monitoring/reporting systems that are suitable for a centralized system. However, many preprogrammed DBMSs for large computers could be used as the basis for an occupational health data system if a centralized configuration were chosen. If microcomputers were used as intelligent front-end processors for such a system, one of the data entry/worksheet programs such as VISICALC could be useful. A considerable amount of programming would be necessary to tie these prepackaged components together; based on the experience of the companies studied, this would probably require two to four programmer/analysts about 10 person-years.

[23] There will no doubt be other users of the system in the Air Force. This study group was unable to estimate the relative costs of these users across system configurations with enough reliability to make comparisons worthwhile.

Telecommunications--Communications in this plan should be through some sort of leased-line arrangement. The amount of communications would not justify a high-speed, hard-wire arrangement; but because of the frequency of communications, the commercial telephone system would be inefficient. The use of a commercial service such as Telenet or Tymnet would probably be best here.

Computer Terminals--There would be one microcomputer system per base that would be configured as an intelligent terminal as well as a stand-alone computer. This machine would consist of a central processor, video display unit, two 100-300-KB floppy disk drives, a 10-MB hard disk unit, and communications interface hardware. Optional equipment would include a dot-matrix printer and graphics display hardware.

COST ESTIMATE: If a large mainframe is available for use as a central data repository, additional disk storage capability will be required to store the occupational health data. Assuming that the system has 117 base-level nodes and that each will require an average of 20 MB of data storage, the system will need an additional 2340 MB. By a rough estimate based on current costs and trends,[24] this will cost approximately $50-$100/MB, or an expense of $117,000 to $234,000. The study group has not determined whether a suitable Air Force computer is already available for this system configuration. If not, one will have to be acquired at an approximate cost of $500,000, including disk storage equipment and system software. Also, a requirement for two additional people to operate and maintain the system is estimated. Assuming a cost of $50,000 per person per year, this would add another $500,000 to the 5-year cost of a centralized system. Also required will be a modern DBMS for the central computer. These vary greatly in cost, but a reasonable estimate would be $1500/month, or $90,000 over the 5-year period.[25] At $10,000 per unit, 117 microcomputer systems would add another $1,170,000. Communications costs--at $8.50/hour, 1 hour/day/node, 260 days/year, and 5 years--would come to $1,292,850.[26] The commercial systems

[24] See, for example, James Martin, Computer Database Organization, p. 4.

[25] Datapro Research Corporation, Buyers Guide to Data Base Management Systems, 1980.

[26] Recall that the estimate for communications with a completely centralized system (that is, communicating with the central computer via dumb terminals rather than microcomputers) was 3 hours/day/base-level system. The estimated threefold increase in communications costs to $3,878,500 over the 5-year period would cost far more than the purchase of the microcomputer equipment.

studied under Task 2 of this project required approximately 10 person-years of effort to develop (exclusive of the development of any data base management utilities which the Air Force should purchase). If five programmer/analysts work for 2 years and one works 3 more years maintaining the system, total applications programmer costs would come to $650,000, assuming a salary plus overhead of $50,000 per year per programmer. The total estimated cost for the centralized system over 5 years should thus be between $3,319,850 and $4,319,850. This would amount to an average of $5,674-$7,384 per base per year.

Distributed System

With this system, the data would be entered and stored through base-level minicomputers. These machines would be linked together in a network that would allow users on one or more central or MAJCOM machines to retrieve data from any base quickly and automatically.

Canned Programs vs Custom Software--The canned program option will only be feasible if extensive use can be made of preprogrammed distributed DBMSs tailored to the minicomputer hardware purchased. The S. C. Johnson OHIS program could be used as part of a distributed system if the base-level computers were DEC VAX machines that could run the MUMPS programs and tie in with other computers through the DECNET distributed network management system. Such machines would be large enough to also support the Uniform Chart of Accounts (UCA) software. Other manufacturers may have systems that support both the MUMPS language and efficient distributed data base management, but it is not recommended that the Air Force take on the task of developing network-management software.

Telecommunications--A distributed processing system such as that described here would require high-speed communications lines connecting the base computers to the central machine. Most distributed data processing systems require more sophisticated equipment than could be provided by ordinary phone lines or AUTOVON.

COST ESTIMATE: A super-minicomputer system will be the central node of the distributed network. The DEC VAX 11/750, which costs about $150,000, is a popular example. Each base-level system would need a compatible but smaller minicomputer. If equipment capable of supporting the MUMPS language and a network management system is acquired for the UCA system, no additional expenses for base-level processors or technical support would be required. Otherwise, another $50,000 per base would be needed (based on the estimated cost of a new VAX machine that will probably be marketed by DEC in mid-1982), or $5,850,000 for all installations. Another $50,000 per year per base (or $29,250,000) will probably also be needed for the technical support staff of these computers. Special hardware and software for the DECNET distributed processing system adds about $7000 to the cost of the 117 base-level systems and the central node, or approximately $826,000. The OHIS software will probably be available in these quantities for a total of $250,000 to $300,000.[27] Personnel requirements will be lower than for the centralized system because most of the software has already been developed; assume two programmer/analysts for 1 year and one for the remaining 4 years, for a total cost of $250,000. The total cost for the distributed system over 5 years is thus apt to be $1,476,000 if base-level computers are available from other Air Force data management programs. If all the hardware for this configuration needed to be purchased, the cost would be approximately $35,100,000. This would amount to an average of $2,523-$60,000 per base per year. These estimates are based on the assumption that Digital Equipment Corporation equipment is acquired, but comparable systems from other vendors are apt to be very similar in price. The $5,850,000 for base-level minicomputers and $29,250,000 for technical support staff under this scheme would presumably be shared by those responsible for the UCA system, which could run on such computers.

Decentralized System

In this system, the data would be entered and stored on base-level computers not directly connected to a central computer.

Canned Programs vs Custom Software--The OHIS system described previously could be acquired for a decentralized system, thus greatly reducing the amount of custom programming required. Whereas the distributed version of the OHIS system would have to run on VAX or comparable super-mini computers, the decentralized version could run on the much smaller DEC PDP 11/23 or 11/34 minicomputers since the base-level machines would not have to communicate directly with one another. Such a system would be sufficiently inexpensive (approximately $30,000/base for hardware and system software in the quantities being considered) that it could be acquired independently of the UCA system. In any event, some modifications to the OHIS software would be required to meet the needs of the Air Force; the extent of these modifications will be determined by this study group after a more detailed examination and evaluation of the actual OHIS code. The modifications would require perhaps 2 person-years of effort.

[27] This should not be taken as a cost quotation from the distributor of the software, only an estimate derived from informal conversations.

Telecommunications--Communications between machines will be sufficiently infrequent that data could be transferred via floppy disks.

COST ESTIMATE: The only hardware requirement for this system would be 117 base-level minicomputers; a typical model would be the DEC PDP 11/23 which would cost an average of $30,000 each (larger bases would need more disk storage capacity), for a total of $3,510,000. As in the distributed system, the OHIS software would come to $250,000-$300,000 total, and the programmer costs would again be estimated at $250,000. Thus, the estimated 5-year cost of the decentralized system envisioned here would be between $4,010,000 and $4,060,000. This would amount to an average of $6,855-$6,940 per base per year.

Concluding Remarks

The distributed system would clearly be the most desirable. It could make use of the OHIS software and would allow easy transfer of data from base to base and to a central computer for epidemiological analyses. While the costs could be high and the technology somewhat uncertain, such a configuration would provide the benefits of decentralization without the disadvantages.

If this path were chosen, considerable coordination between those responsible for acquiring the UCA system and the USAF Occupational Health Program would be necessary to ensure that the system could support the MUMPS language and distributed processing software, neither of which were envisioned when the specifications for the UCA system were written. If the Occupational Health and UCA systems could run on the same computers, however, the benefits of the distributed system might well be obtained at a considerably lower marginal cost. If these arrangements can be made, this study group recommends the distributed system option; if not, the cost would be prohibitive.

The partially centralized system could be the least expensive. The disadvantages of such a system stem mainly from the fact that no acceptable prepackaged occupational health data management system is available for a centralized operation on the scale required by the Air Force, so considerable time and effort would be required to actually implement this system. As noted, the Air Force should strongly resist the temptation to hire many programmers in an attempt to shorten the actual time needed to design and code the system. Based on the experience of many in industry, that tactic is unlikely to be successful. Because the time required to actually implement this system is double that of the other options, and because numerous uncertainties are involved in the production of large software packages, the partially centralized option is the least preferred of the three.

Considering both cost and convenience factors, the decentralized system configuration is the most cost-effective, assuming that the OHIS software is procured but cannot be run on the UCA base computers. Both the cost and time estimates would increase considerably if a decentralized custom-programmed system were envisioned. The cost estimates above are relatively "hard" except for the estimates of the amount of programmer time needed to modify the system to Air Force requirements. These estimates will be revised when this study group examines the OHIS system more closely.

The following table summarizes the three most promising system configurations and the relative cost estimates. The lower cost is essentially the marginal cost of adding the occupational health software to existing mainframes or the UCA minicomputers, assuming this is possible. The higher estimate includes the cost of purchasing such hardware, although these costs might ultimately be justified across organizations. The higher estimate also includes the high-end estimate of other costs when only a range could be determined.

SUMMARY OF THREE POSSIBLE SYSTEM CONFIGURATIONS

Characteristic	System Configuration		
	Partly Centralized	Distributed	Decentralized
Central computer (memory/disk)	Mainframe (10 MB/2500 MB)	Supermini (1 MB/100 MB)	
Base computer (memory/disk)	Micro (64 KB,10 MB)	Mini (1 MB/20 MB)	Mini (128 KB/20 MB)
Data communications	Digital network	Digital network	Floppy disk
Software available (supplier)	DBMS only (several suppliers)	Complete system (S. C. Johnson)	Complete system (S. C. Johnson)
Development language	Pascal/Ada	MUMPS	MUMPS
Development time	2 years	1 year	1 year
Development programmers	Five	One	One
5-year cost (millions)	$3.2 - $4.2	$1.5 - $35.1	$4.0 - $4.1
Cost/base/year (thousands)	$5.7 - $7.4	$2.5 - $60.0	$6.9
Relationship to other systems	Shares mainframe with other users	Uses same base computers as UCA	Independent
Evaluation	3rd choice	1st choice if UCA-compatible	2nd choice

FILMED

-8